PERGAMON

Pattern Recognition 33 (2000) 385–398

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Morphological waveform coding for writer identification

E.N. Zois, V. Anastassopoulos*

Electronics Laboratory, Physics Department, University of Patras, Patras 26500, Greece

Received 23 April 1997; received in revised form 25 August 1998; accepted 16 February 1999

Abstract

Writer identification is carried out using handwritten text. The feature vector is derived by means of morphologically processing the horizontal profiles (projection functions) of the words. The projections are derived and processed in segments in order to increase the discrimination efficiency of the feature vector. Extensive study of the statistical properties of the feature space is provided. Both Bayesian classifiers and neural networks are employed to test the efficiency of the proposed feature. The achieved identification success using a long word exceeds 95%.

Keywords: Writer identification; Person verification; Morphological features; Waveform coding

1. Introduction

Handwritten patterns constitute the behavioral part of biometrics approach towards person verification which is not invasive in contrast to that of physiological biometrics (fingerprints or iris characteristics). Off-line writer verification systems based on signatures have been studied extensively in the past [1]. A writer verification system based on handwritten text is expected to provide discrimination results equivalent to those obtained from signatures, since text has been reported to comprise rich and stable information [2]. Furthermore, a handwritten sentence can be determined and changed by the writer at will. In high security data systems like those involved in financial transactions, the first step towards reaching a specific person's data is usually carried out by means of the personal identification number (PIN) number. However, handwritten patterns such as the signature or a word can be used on a complementary basis to improve system reliability. In order to increase further the reliability of the verification system, many handwritten words can be used by means of fusion techniques [3].

In general, writer discrimination and verification approaches based on handwritten text are hardly found in the literature [4,5]. Security reasons or specific law restrictions have prevented serious results of significant importance on the topic from publicity [6]. To the knowledge of the authors few publications are related to writer discrimination and especially to feature extraction [7,8]. Feature extraction from handwritten text can be carried out using approaches that resemble those of signature verification. However, features which contain information of the trace of each word are usually preferable.

In this work a writer identification method is proposed, which is based on the use of a single word. The image of the word is properly preprocessed and projected onto the horizontal direction. Projection functions have been used widely in the literature for contour feature extraction [9,10], signature analysis [11,12] and recognition of handwritten characters (Latin, Chinese, etc.) and numerals [13,14]. A projection is a global shape descriptor which provides a kind of line image coding [10]. The obtained projections are segmented in parts which are morphologically [15,16] processed in order to obtain the required feature vector. The morphological processing is a type of granulometry, i.e. the measure of area reduction through successive openings [16,17]. Two different types

* Corresponding author. Tel.: + 30-61-996-147; fax: + 30-61-997-456.

E-mail address: vassilis@physics.upatras.gr (V. Anastassopoulos)

0031-3203/00/$20.00
PII: S0031-3203(99)00063-1

of windows are applied on the segments of the projection functions to control the flow of information from one part to the other. The blanks between the letters are also considered in the formation of the feature vector.

Both the statistical properties of the feature space and the capability of the specific features for writer identification are extensively studied. This study includes the underlying pdf for the feature vector components as well as the separability of the clusters in the feature space. For this purpose a cluster separability measure is proposed and analyzed. Next, two different classification schemes are tested. Namely, the Bayesian classifier and the neural networks. In the classification procedure, the binary decision problem (writer verification: is this person he who claims to be?) and the general classification problem (writer identification: identify a writer among many others) are studied. A writer verification error smaller than 5% is achieved. The error becomes smaller while increasing feature dimensionality. A database [18] was created employing 50 writers, while an English and the equivalent Greek word were used to demonstrate that the method is language independent.

The paper is organized as follows. In Section 2 the developed database is presented. In Section 3 the procedure used for feature extraction is analyzed. In Section 4 the formation of the feature space is explained and criteria for measuring cluster intra-distance and inter-distance are presented. An extensive study on the statistics of the feature space is also carried out. Section 5 deals with the experimental performance of two classification schemes in the multicategory and the two-category case. The conclusions are drawn in Section 6.

2. The database and data preprocessing

Data acquisition and preprocessing constitute an essential step towards feature extraction and writer discrimination. Specifically, the acquisition stage affects the quality of the image, which in turn determines the reliability of the feature vector and the recognition procedure. The off-line procedures dealt in this work give full discretion to obtain good quality images.

2.1. The database

The database employed for writer identification was created so that two different issues are appropriately addressed. Firstly, a considerable number of samples was recorded to ensure the validity of the experimental results. Secondly, the database was constructed using both an English and a Greek word in order to show that the applicability of the feature vector is independent of the language used. Accordingly, a blank page of size A4 is divided in 45 shells (15 lines × three columns). Each writer had to fill in all the shells of the page with the word

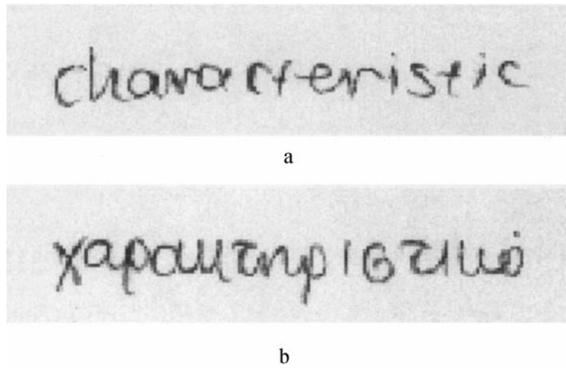

Fig. 1. Scanned word samples from the database: (a) English; (b) Greek word.

'characteristic' and the shells of another page with the equivalent Greek word (Fig. 1). The only constraint was that the writer should write down the words inside the shells. Fig. 1 shows a word sample by a specific writer in a specific shell after the scanning process. For each shell an image file with dimensions 230 by 70 pixels is created, with 256 gray levels. A total of 50 writers has been recorded in the database which is available to the research community through Internet [18].

2.2. Preprocessing

The database is firstly preprocessed so that the derived features are as far as possible independent of the writing conditions. Preprocessing is employed to eliminate redundant information caused by the randomness of scanning, the difference between the pens used by different writers as well as the capability of the paper used to soak up ink. In all the above cases the use of the most appropriate image enhancement techniques depend on operators experience. The preprocessing algorithms applied in this work are image thresholding and curve thinning. Due to satisfactory image acquisition conditions, these two algorithms are regarded adequate to reveal the special characteristics of each writers handwriting.

Firstly, thresholding was applied to both English and Greek words of the database. Histogram thresholding is used in order to separate gray (word) and white pixels. The threshold used was between 170 and 180 gray value for all images. A thresholded result (black and white image) is shown in Fig. 2a accompanied by its thinned version (Fig. 2b). The thinning process produces the trace of the thresholded images with only one pixel width. The algorithm realizes simple morphological transformations (openings) with four structuring elements of different orientations applied successively on the image only once.

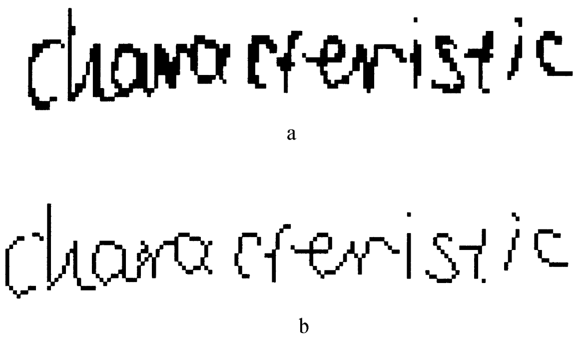

Fig. 2. Preprocessing stage: (a) thresholded image; (b) thinned image.

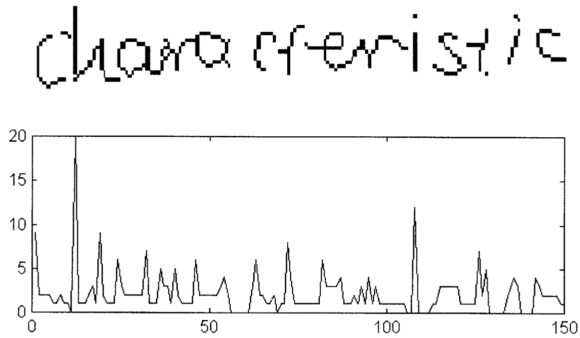

Fig. 3. The projection function corresponding to the image of the word 'characteristic'.

3. Feature extraction

The feature extraction procedure is described in this section. The proposed feature vector is obtained by means of morphologically transforming the projection functions of the thinned images. The length of the projections is firstly normalized. Afterwards, morphological openings are applied to the segments of the projection for feature extraction.

3.1. The normalized projection function

The projection function is derived by mapping the two-dimensional thinned image to a one-dimensional function. This function contains information about the spatial pixel distribution of the word trace along the horizontal direction. More specifically, it is formed by measuring the black pixels contained in each column of the thinned image [9–14]. In Fig. 3 the thinned image as well as the corresponding projection function $f(x)$ are shown. The zeros of the function which correspond to the blanks between the letters contain significant information about the specific handwriting. Accordingly, two versions of the projection function are created as shown in Fig. 4. Both functions have been shifted to the origin, whereas in the second function (Fig. 4b) the zero bins are eliminated. Furthermore, the length of the function is not constant even for those samples written by the same writer. In order to make the feature independent of the word length the functions are resampled so that the total number of bins is 100, as shown in Fig. 5. Resampling to a constant length of 100 points incorporates antialising procedures and is carried out using a special MATLAB routine. Both functions are invariant under translation since they are shifted to the origin. Rotation invariance is assumed to exist since each person writes along the horizontal line. Hereafter, the normalized in length function containing the blanks will be addressed as the

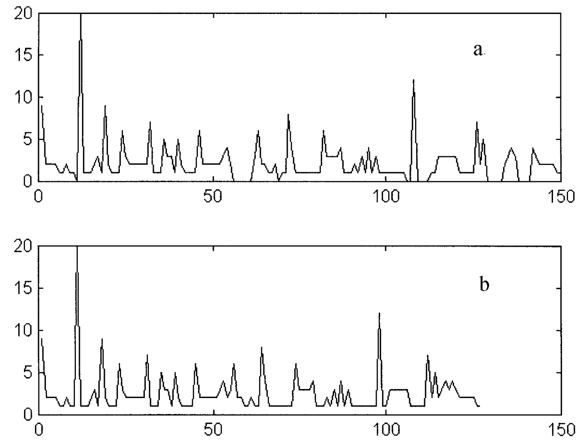

Fig. 4. Raw projection functions: (a) function containing blanks between letters; (b) function without blanks.

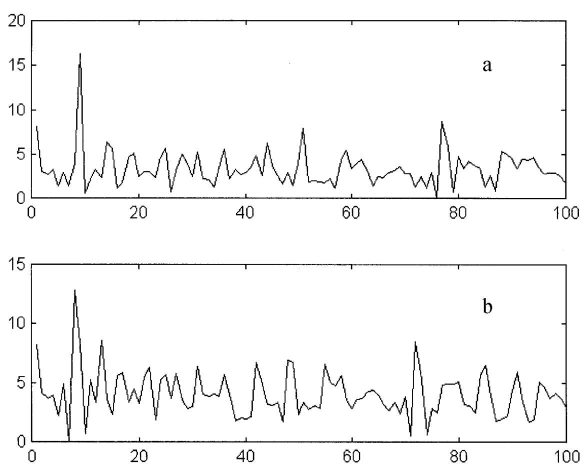

Fig. 5. Final projection function resampled to 100 bins: (a) projection function (PF); (b) compressed projection function (CPF).

'projection function (PF)' while the normalized in length function without blanks as 'compressed projection function (CPF)'.

The resulting projection functions are of too large dimensionality to be considered as features for discrimination. It is well known that large dimensionality is the curse of every pattern recognition technique. In this work, the information content of both the PF and CPF is further compressed by means of morphological transformations.

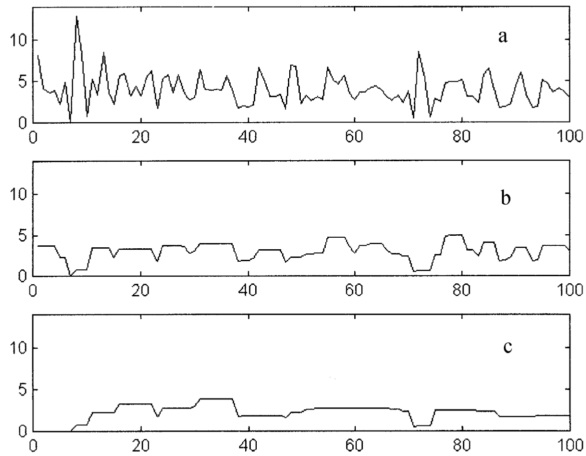

Fig. 6. The morphological opening on the projection function: (a) initial projection function; (b) opening with line SE of length 3; (c) opening with line SE of length 7.

3.2. Morphological transformation and the feature vector

Mathematical morphology [15] is based on set theory and deals with the interaction of one set called the structuring element (SE) with the set to be transformed. Information on the characteristics of a binary object can be obtained by various transformations with different SE [16,17]. The two basic operations in morphology are erosion and dilation. Erosion ($\ominus$) of a set X by a SE B is a shrinking operation defined as follows [15]:

$$X \ominus B = \bigcap_{b \in B} X_{-b}, \tag{1}$$

while dilation ($\oplus$) is an expanding operation

$$X \oplus B = \bigcup_{b \in B} X_{+b}, \tag{2}$$

where the subscript denotes geometric translation. On the other hand, the morphological opening ($\circ$) is defined as an erosion followed by a dilation with the same SE

$$X \circ B = (X \ominus B) \oplus B. \tag{3}$$

After an opening operation the original object X has been smoothed from those details which the structuring element B does not fit in [17]. This fact can be used to measure the loss of information when gradually increasing the size of the structuring element. In case of one-dimensional function, morphological transformations operate on the umbra [15] of the function $f(x)$ regarding it as a set. Assuming that the SE $g(x)$ is a line segment with length L and zero valued in the domain it is defined, erosion and dilation are expressed respectively as

$$(f \ominus g)(x) = \min_{\substack{z \in D \\ z-x \in G}} \{f(z) - g(z-x)\} = \min_{\substack{z \in D \\ z-x \in G}} (f(z)) \tag{4}$$

and

$$(f \oplus g)(x) = \max_{\substack{z \in D \\ z-x \in G}} \{f(z) + g(z-x)\} = \max_{\substack{z \in D \\ z-x \in G}} (f(z)) \tag{5}$$

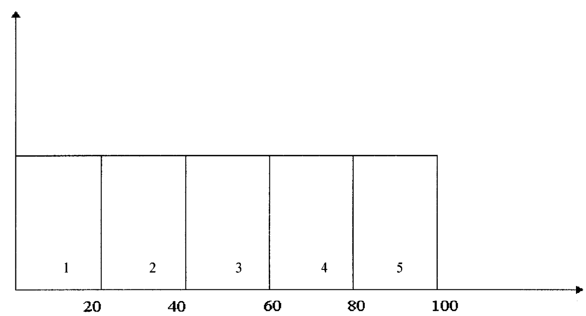

Fig. 7. Partition of the region of the projection function $f(x)$ for feature vector extraction. Rectangular windows. The number of segments is 5.

where D is the domain of $f(x)$ and G is the domain of $g(x)$. The morphological opening $f \circ g$ of the function f by the SE g is defined according to (3). The measurement of the gradual reduction in the area using openings is called granulometry or pattern spectrum [16,17] and is used, in this work, to give the feature vector for writer discrimination. The differences between successive openings denote the amount of information that is removed by the increasing in size structuring element. In Fig. 6 are shown graphically the results when successive openings are applied to the PF $f(x)$ with line SE $g(x)$ of length $L = 3,7$, respectively.

The final feature vector is created merely by partitioning the projections into a number of segments and measuring the relative amount of area that the two SE (with lengths 3 and 7) reject in each block. Fig. 7 shows the case of segmenting the region of $f(x)$ into five sub-blocks, thus extracting a 10-dimensional feature vector.

Generally, the components of the feature vector p are defined as follows

$$p_i = \left[\frac{\mathrm{Mes}(f) - \mathrm{Mes}(f \circ g_3)}{\mathrm{Mes}(f)}\right]_i \quad \text{for } i = 1, \ldots, m \qquad (6)$$

and

$$p_{i+m} = \left[\frac{\mathrm{Mes}(f \circ g_3) - \mathrm{Mes}(f \circ g_7)}{\mathrm{Mes}(f)}\right]_i \quad \text{for } i = 1, \ldots, m \qquad (7)$$

where g_3 and g_7 denote the line SE with length $L = 3$ and 7 respectively, Mes(.) is the area under the function in the argument, i stands for the ith segment of the region of the function $f(x)$ and m is the partition cardinality. The feature vector p is similar to the pattern spectrum described in the literature [16,17]. Each component p_i of the feature vector p describes the pixels allocation along the trace of the word in each segment. Actually, p_i describes the fine details of the trace while p_{i+m} describes the coarse distribution of the pixels in each segment. Specifically, if the first m p_i's are large enough they reveal a tendency of the writer to persist in vertical lines. Whereas, if the rest p_i's are quite large smoother distribution of the lines into both directions is expected.

The feature vector components corresponding to the same segment (i.e. p_i and p_{i+m}) are somewhat correlated. This will be examined in the next section where covariance matrix properties are analyzed. Furthermore, some kind of correlation is expected between neighboring components p_i and p_{i+1}. This is due to the fact that information near the edge of the segment is prone to moving towards the next segment owing to both handwriting variations and the resampling procedure. Partitioning the functions with overlapping trapezoidal segments a new feature vector q is derived which is more stable to information shift. Fig. 8 shows the way the above idea is implemented in case of partitioning into five segments. Specifically, for the evaluation of the area of a function (Mes(.)) in a specific segment, part of the area of the adjacent segment is considered multiplying by a trapezoidal instead of a rectangular window. The feature vectors p and q have been extensively tested as far as the achieved separability in the feature space is concerned.

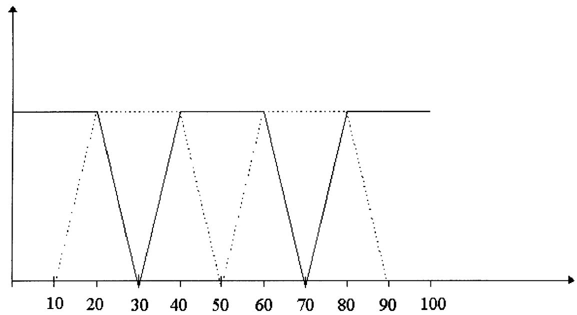

Fig. 8. Partition of the region of the projection function $f(x)$ for feature vector extraction. Five trapezoidal windows are used.

4. Feature space statistics and properties

The statistical characteristics of the derived four types of features are exploited in this section and conclusions are drawn about their classification capabilities. The extent of the clusters into the feature space is examined by means of the eigenvalues of the cluster covariance matrices. Information on the correlation of the features can also be obtained from these covariance matrices. Next, the pdf of the features is exploited using the K–S fit test. The Gaussian pdf is found to be a good candidate for one of the types of the features employed. For the same feature, maximum cluster separability in the feature space is observed. A cluster separability measure is introduced and analyzed theoretically. This measure is used to assess writer separability.

4.1. Class covariance matrices

The covariance matrix of a population is a means of measuring the variance of each component in the feature space as well as the correlation between the feature components. It also provides, in case we are confronted with high dimensionality data sets, a measure for the intrinsic dimensionality. Thus, diagonalization procedures provide the principal components of the orthogonal features which lie along the eigenvectors of the covariance matrix. The ability of the above procedure to make apparent the most dominant components, leads to reduction of the original feature space.

Evaluating the covariances for every feature type and for a partition range of 5–10 segments, the maximum number of dominant eigenvalues is found to be smaller than eight. The small intrinsic dimensionality of the feature space results in the following. Firstly, the number of samples per writer is considered adequate for feature vector mean and covariance estimation with reduced bias, thus giving consistent error probabilities. Secondly, the use of distance as a similarity measure, which is a mapping to a one-dimensional space, causes small distortion to the classification information since the intrinsic dimensionality of the original space is small. The eigenvalues of the covariance matrix corresponding to the first writer and for a specific partition level are given in Table 1. Additionally, working out the covariances of all data sets it was found that the eigenvectors corresponding to the minimum eigenvalues differ from writer to writer. Consequently, there is not a common base in the feature space which could be used for simultaneously rotating all clusters in order to reduce the dimensionality of the feature space and decorrelate the features.

Table 1
Eigenvalues of the covariance matrix for writer No.1. The features were extracted using five trapezoidal windows

0.0003	0.0005	0.0006	0.0018	0.0032	0.0036	0.0048	0.0069	0.0091	0.0134

Table 2
Correlation coefficients for features from trapezoidal windows without blanks and compressed projection function

1.0000	0.1078	0.3882	0.3409	− 0.3434	0.1286	− 0.3043	0.0492	0.1060	0.2510
0.1078	1.0000	− 0.0049	0.5087	− 0.1226	0.0481	− 0.3393	0.0912	− 0.0493	− 0.0959
0.3882	− 0.0049	1.0000	0.1224	0.2385	0.1026	0.0099	− 0.1267	− 0.1192	0.0631
0.3409	0.5087	0.1224	1.0000	− 0.0427	0.6019	− 0.0695	0.2438	0.2664	− 0.0951
− 0.3434	− 0.1226	0.2385	− 0.0427	1.0000	− 0.0825	0.7678	− 0.2553	0.2318	− 0.1416
0.1286	0.0481	0.1026	0.6019	− 0.0825	1.0000	0.0570	0.4989	0.1640	0.0408
− 0.3043	− 0.3393	0.0099	− 0.0695	0.7678	0.0570	1.0000	− 0.3006	0.3297	− 0.0867
0.0492	0.0912	− 0.1267	0.2438	− 0.2553	0.4989	− 0.3006	1.0000	0.0556	0.4705
0.1060	− 0.0493	− 0.1192	0.2664	0.2318	0.1640	0.3297	0.0556	1.0000	− 0.2697
0.2510	− 0.0959	0.0631	− 0.0951	− 0.1416	0.0408	− 0.0867	0.4705	− 0.2697	1.0000

A measure of feature correlation can be obtained by examining the non-diagonal elements of the covariance matrix. If we define the covariance matrix of a data cluster C_i and its elements c_{ij} as

$$c_{ij} = E\{(x_i - m)(x_j - m)^t\} \tag{8}$$

where x and m are the feature vector and the sample mean, respectively, for a specific writer, then the components r_{ij} (correlation coefficients) of the correlation matrix R emanate from the relation

$$r_{ij} = \frac{c_{ij}}{\sqrt{c_{ii}\, c_{jj}}}. \tag{9}$$

The above correlation coefficients were evaluated for the features extracted using rectangular and trapezoidal windows. Next, the same calculations were carried out for both the PF and the CPF, respectively. It is concluded that features are less correlated when rectangular windows are employed. This result was expected since using trapezoidal windows, neighboring feature components (e.g. p_3 and p_4) share a common amount of information. However, even in the case of trapezoidal windows the correlation of the features seldom exceeds 0.3, a value which can be considered to represent weak correlation. Finally, correlation coefficients remain almost the same when the features are coming from functions with or without blanks between letters. Tables 2 and 3 show the correlation coefficients between rectangular and trapezoidal windows for one writer.

4.2. Statistical behavior of the proposed features

The statistics of the feature components and their agreement with the normal density is examined by means of the statistical fit tests [20], and especially the Kolmogorov–Smirnov (K–S) test [21]. Specifically, we have to disprove, to a certain required level of significance, the null hypothesis H_0 that a data set follows a predetermined distribution function. Disproving the null hypothesis in effect we prove that the data set comes from a different distribution. On the other hand, proving the null hypothesis shows that the data set is consistent with the considered distribution function. Other methods found in the literature for inspecting the form of a distribution are the Parzen windows and the K-nearest neighbors [22].

In order to examine the similarity between two cumulative functions we define D as the maximum observed value of their absolute difference as shown in Fig. 9:

$$D = \max_{-\infty < x < \infty} |S_N(x) - P(x)| \tag{10}$$

where $S_N(x)$ is the cumulative function of the sample data and $P(x)$ is a known distribution function. Under certain conditions and given that hypothesis H_0 is true, the Kolmogorov–Smirnov statistic D follows the cumulative distribution [21]

$$F_D(D) = 1 - 2\sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2\lambda^2}, \tag{11}$$

where $\lambda = (\sqrt{N} + 0.12 + 0.11/\sqrt{N})D$ and N is the number of data samples used. We must reject H_0 if D is larger than a constant c. This constant is determined in terms of the *significance level* α [20]:

$$\begin{aligned}\alpha(c) &= P\{D > c|H_0\}\\ &= 1 - P\{D < c|H_0\}\\ &= 1 - F_D(c)\end{aligned} \tag{12}$$

Table 3
Correlation coefficients for features emerging from rectangular segments without blanks and compressed projection function

1.0000	0.1178	0.1286	0.2756	− 0.2234	0.2001	− 0.3361	− 0.0701	0.1524	0.2739
0.1178	1.0000	− 0.1017	0.2624	− 0.0314	0.0387	− 0.3671	0.1033	0.0414	− 0.1651
0.1286	− 0.1017	1.0000	− 0.0616	− 0.0765	0.1204	− 0.0593	− 0.0772	− 0.1302	0.0528
0.2756	0.2624	− 0.0616	1.0000	0.0510	0.2293	− 0.0198	0.0905	0.5343	− 0.3166
− 0.2234	− 0.0314	− 0.0765	0.0510	1.0000	− 0.1891	0.4613	0.1166	0.3905	− 0.1579
0.2001	0.0387	0.1204	0.2293	− 0.1891	1.0000	0.1450	0.1656	0.0887	0.1128
− 0.3361	− 0.3671	− 0.0593	− 0.0198	0.4613	0.1450	1.0000	− 0.3532	− 0.0082	0.0637
− 0.0701	0.1033	− 0.0772	0.0905	0.1166	0.1656	− 0.3532	1.0000	0.2982	0.0383
0.1524	0.0414	− 0.1302	0.5343	0.3905	0.0887	− 0.0082	0.2982	1.0000	− 0.3551
0.2739	− 0.1651	0.0528	− 0.3166	− 0.1579	0.1128	0.0637	0.0383	− 0.3551	1.0000

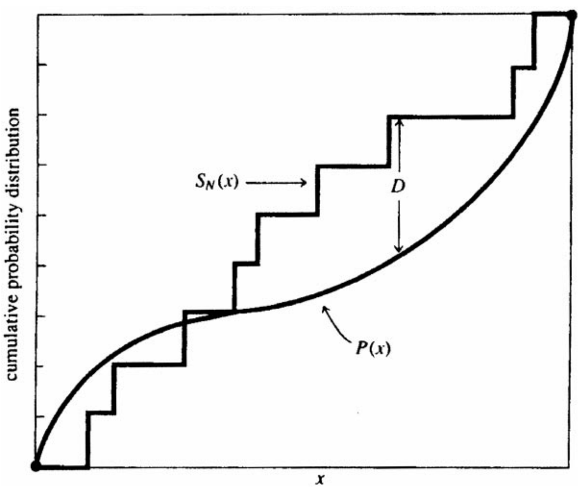

Fig. 9. Kolmogorov–Smirnov statistic D.

From Eq. (12) we can determine c for a specific significance level α. Accordingly, $F_D(c) = 1 - \alpha$ verifies our confidence about the validity of H_0. In practice, for an observed value D_1 this confidence is expressed as the probability D_1 can be of the smallest values of D, i.e.

$$P(D_1 < D) = 1 - P(D < D_1) = 1 - F_D(D_1) \quad (13)$$

In our experiment a data set of $n = 45$ points (number of words) was available for each writer and each feature component for the K–S test to be applied. The value of D as well as the confidence about H_0 were evaluated for 500 data sets (50 writers × 10 feature vectors) and for a low-level partition (five segments). Fig. 10a provides the histogram of the measured values of D, while in Fig. 10b the distribution of our confidence about H_0 is shown. The majority of the D's is around 0.1 which corresponds to a degree of confidence larger than 75%. Similar results were obtained using the Greek word. This supports our claim that for each individual feature component and for each writer the normal density can be considered as good approximation to the data.

It is worth mentioning that the best approximation to Gaussian statistics was achieved by the features derived using the compressed projection functions and trapezoidal windows in the feature extraction procedure. Similar experimental procedure was carried out for features obtained using rectangular windows and/or the simple projection function. Normal pdf hypothesis was not found strong enough in these cases. Hence, the trapezoidal windows are proved to be a natural process which strengthens the validity of Gaussian pdf for the derived features.

4.3. Cluster separability measure

A person's handwriting is not precisely repeatable since it changes with physical and mental state, as well as with the age [1]. Generally, we can distinguish between two kinds of handwriting variability. The intraclass variability which describes the variations within a class (same writer), and the interclass variability which describes the differences between writers. Ideally, intraclass variability should be as low as possible, while interclass variability should be as large as possible. In practice, classes are not well separated. A quality factor which indicates the separability between two classes is introduced here. This quality factor expresses the maximum theoretical error in classifying the samples of two clusters when these two clusters are normally distributed and intermixed across the line of their larger variances.

Let us consider two normally distributed populations ω_a and ω_b with two dimensional pdf's $p_a = N(\mu_{ax}, \sigma_{ax}, \mu_{ay}, \sigma_{ay})$ and $p_b = N(\mu_{bx}, \sigma_{bx}, \mu_{by}, \sigma_{by})$, respectively, and the same a priori probabilities $P(\omega_a) = P(\omega_b)$. The above example can easily be generalized for higher dimensions where real problems are met. Fig. 11 shows the contour plot of two normal distributions having variances in x and y axes equal to unity, while their means are selected so that their Euclidean distance

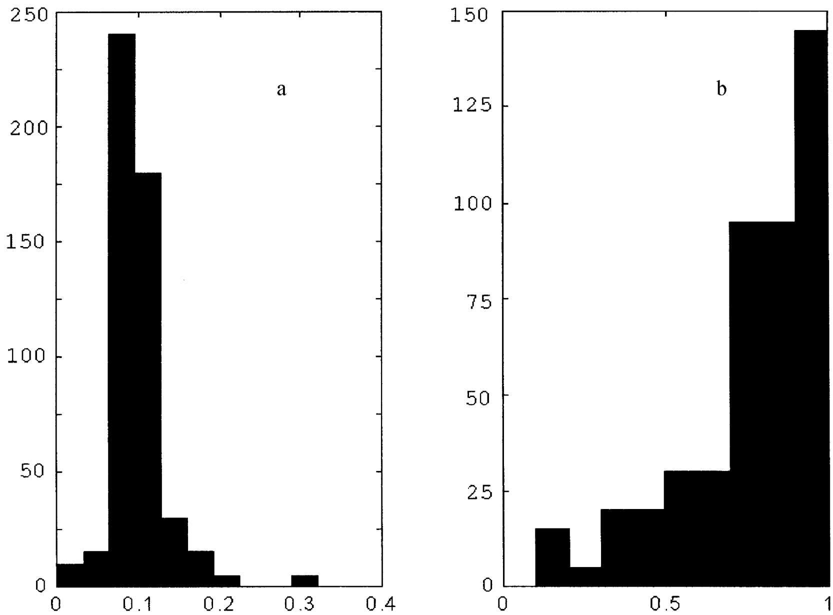

Fig. 10. (a) Histogram of D for 500 data sets (50 writers and a 10-dimensional feature vector) using the compressed projection function and trapezoidal windows. Partition level is 5 and the word used is the English one. (b) The corresponding histogram for the degree of confidence $(1 - F_D(D))$ about the hypothesis H_0 for 500 data sets.

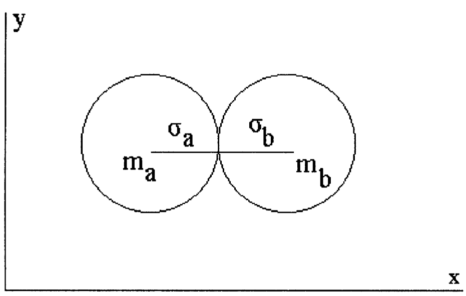

Fig. 11. Contours corresponding to $\sigma_i = 1$, for two normal distributions with equal variances and mean distance equal two.

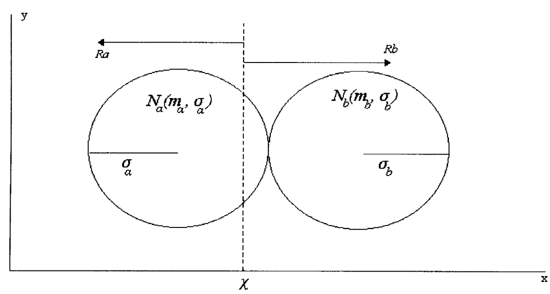

Fig. 12. Decision boundary for the two class problem.

equals two. Inside each individual circle lies 60% of the samples.

We seek a quantity which would express the separation between these two populations. This separation is easier when the populations are quite distant and the dispersion of each one is small. In order to measure the distance between two populations, we use the Euclidean distance of their means, whereas the dispersion of each population is measured using the largest eigenvalue λ_{imax} of its covariance matrix. This eigenvalue is related to the length of the largest semiaxis of the cluster hyperellipsoid and the corresponding standard deviation $\sqrt{\lambda_{imax}} = \sigma_{imax}$. Thus, we introduce the ratio R in order to express the separability of two classes as follows:

$$R(a, b) = \frac{D(m_a, m_b)}{\mathrm{Max}(\sqrt{\lambda_a}) + \mathrm{Max}(\sqrt{\lambda_b})}, \tag{14}$$

where $D(m_a, m_b)$ is the Euclidean distance of the population means and $\mathrm{Max}(\lambda)$ is the maximum eigenvalue for each class. In the worst-case scenario (maximum classification error), the eigenvectors of the covariance matrix which correspond to the greatest eigenvalues lie in the direction $\{\bar{m}_a - \bar{m}_b\}$ as shown in Fig. 12. The separability ratio R in this case is theoretically calculated as

$$R_T = \frac{D(m_a, m_b)}{\mathrm{Max}(\sqrt{\lambda_a}) + \mathrm{Max}(\sqrt{\lambda_b})} = \frac{m_b - m_a}{\sigma_a + \sigma_b} \tag{15}$$

and the maximum theoretical error when the above situation holds is taken from

$$\begin{aligned}P(error) &= P(x, y \in R_b|\omega_a) + P(x, y \in R_a|\omega_b)\\ &= \int_{R_b} p(x, y|\omega_a)\, P(\omega_a)\, \mathrm{d}x\, \mathrm{d}y\\ &\quad + \int_{R_a} p(x, y|\omega_b)\, P(\omega_b)\, \mathrm{d}x\, \mathrm{d}y\\ &= \int_{y=-\infty}^{\infty}\int_{x=\chi}^{\infty} p_a(x, y \mid \omega_a)\, P(\omega_a)\, \mathrm{d}x\, \mathrm{d}y\\ &\quad + \int_{y=-\infty}^{\infty}\int_{x=-\infty}^{\chi} p_b(x, y|\omega_b)\, P(\omega_b)\, \mathrm{d}x\, \mathrm{d}y, \end{aligned} \quad (16)$$

where the a priori probabilities $P(\omega_a)$ and $P(\omega_b)$ are considered equal to 0.5 and the bivariate densities p_a and p_b are decorrelated. This way p_a and p_b are separable with respect to x and y so that Eq. (16) becomes

$$\begin{aligned}P(error) = \frac{1}{2}\Bigg[&\frac{1}{\sqrt{2\pi}\sigma_{ax}}\int_{x=\chi}^{\infty} \exp\left(-\frac{(x-m_{ax})^2}{2\sigma_{ax}^2}\right)\mathrm{d}x\\ &+\frac{1}{\sqrt{2\pi}\sigma_{bx}}\int_{x=-\infty}^{\chi} \exp\left(-\frac{(x-m_{bx})^2}{2\sigma_{bx}^2}\right)\mathrm{d}x\Bigg].\end{aligned} \quad (17)$$

This integral becomes minimum for $\sigma_{ax} = \sigma_{bx}$ when χ is placed in the middle of $m_{bx} - m_{ax}$ [22]. In this case, χ equals $(m_{ax} + m_{bx})/2$ and using Eqs. (15) and (17) becomes after some mathematical manipulations [24]

$$P(error) = 0.5 - 0.5\mathrm{erf}\left(\frac{R}{\sqrt{2}}\right). \quad (18)$$

Using simple MATLAB routines the theoretical maximum classification error is found to be 16.5% for $R = 1$ which is the case described in Fig. 11. For $R > 1$ or in case that the clusters higher dispersion is not in the $\bar{m}_a - \bar{m}_b$ direction, the success is expected higher.

Experimentally, the proposed quality factor R was evaluated for each feature type, considering ω_1 as the writer under examination and ω_2 as the set of all the other writers (totally 49). The ω_2 class can be viewed as noise which must be rejected from the genuine classes ω_1. A low-level partition was used (five segments) for the projection functions. The results are shown in Fig. 13. From this figure it is obvious that the highest values of R (higher separability) are obtained when trapezoidal windows and compressed projection functions are employed (Fig. 13a). Therefore, the experimental classification procedure in the next section is carried out using only the corresponding type of features.

5. Classification approaches and discrimination results

In order to evaluate the performance of the proposed features for writer discrimination, a comparative study is

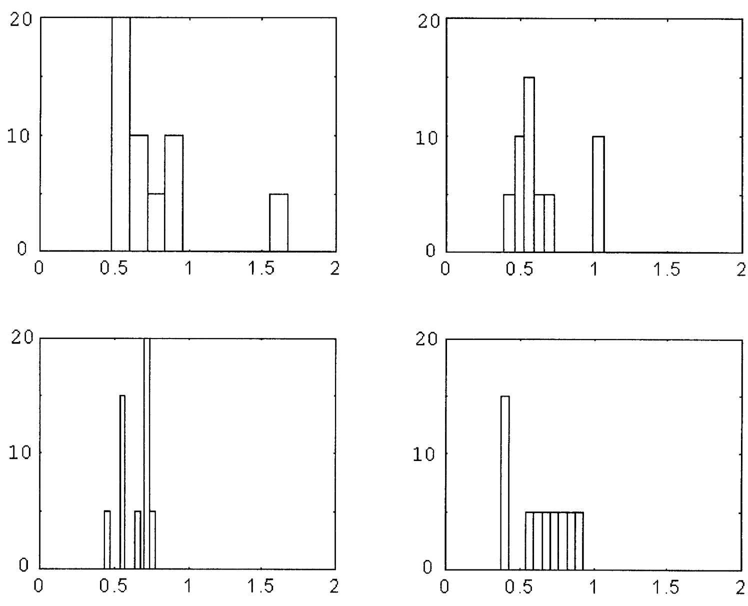

Fig. 13. Histogram of R for all pairs ω_1 and ω_2 obtained from 50 writers and the English word. Partition level is 5 (10-dimensional feature vector). (a) Features from compressed projection functions and trapezoidal windows. (b) Features from compressed projection functions and rectangular windows. (c) Features from simple projection functions and rectangular windows. (d) Features from simple projection functions and trapezoidal windows.

carried out by means of two well-established classification schemes. The conventional Bayesian approach using weighted distance measures is examined first. The simple multilayer perceptron is tested, next.

In biometrics the most common issues concerning the effectiveness of features, which potentially describe the behavior of a person, are identification and verification. In the first case, an unknown sample is classified among a number of writers thus answering the question "*who does this sample belong to*?". The latter case deals with the problem of deciding whether a word or text belongs to a specific writer or not. So, the question arising here is "*does that sample belong to that specific person*?". In both cases, the classification approach followed was to employ the database described in Section 2 and form the feature space and the corresponding covariance matrices. The word under test is assigned to a specific writer based on the distance of the feature vector from the corresponding cluster center.

5.1. Classification using Bayesian approach

According to the material exposed in the previous section, the Gaussian pdf is a satisfactory approximation regarding one set of the features. For this same feature, the maximum separability in the feature space is achieved based on the ratio R. In this subsection, the identification problem is considered using this particular feature. The classification criterion is based on the weighted distance from the center of each cluster

$$d_i = (x - m_i)^t C_i^{-1}(x - m_i), \tag{19}$$

where m_i is the center of each cluster, and C_i is the corresponding covariance matrix. In case that each C_i equals I, the Euclidean distance is obtained. The estimation of m_i and C_i constitutes the training procedure of the conventional Bayesian classifier.

For the 50-writer group, the classification procedure was based on the leave-one-out-method [22]. According to this method the covariance matrix C_i as well as the center m_i of each cluster in the feature space are found using 2249 points. Then the remaining point is assigned to the writer with the minimum d_i. The procedure is repeated for all 2250 points of the feature space. Both the mean and the covariance must be determined every single time. However, methods have been developed and used here that overcome the problem of designing $N = 2250$ classifiers [23]. This is done by properly weighing the mean and the covariance of each class using the following relations:

$$\hat{M}_{ik} = \hat{M}_i - \frac{1}{N_i - 1}(X_k^{(i)} - \hat{M}_i), \tag{20}$$

$$\hat{\Sigma}_{ik} = \hat{\Sigma}_i + \frac{1}{N_i - 2}\hat{\Sigma}_i - \frac{N_i}{(N_i-1)(N_i-2)}(X_k^{(i)} - \hat{M}_i)(X_k^{(i)} - \hat{M}_i)^t, \tag{21}$$

where $\hat{M}_{ik}$, $\hat{\Sigma}_{ik}$ are the mean and covariance estimates of the i-class without the $X_k^{(i)}$ sample, N_i is the class population, and $\hat{M}_i$, $\hat{\Sigma}_i$ are the estimates when all the samples are used. The method was applied to both the English and the Greek word. The partitions (number of segments) employed in order to examine the efficiency of the feature vector were 5, 6, 7, 8, 10, 12 and 15 leading the dimensionality of the feature space to 10, 12, 14, 16, 20, 24, and 30, respectively.

The experimental results of the above 50-writer identification procedure are presented in Table 4a and b. The total identification success was found 92.48% for the English word and 92.63% for the Greek one. This is considered satisfactory for simultaneously discriminating 50 writers and using only one word. It has been observed that beyond a partition level (which in our case is 10) the feature efficiency drops drastically. This is due to the fact that there exist highly correlated feature components, a large number of which is zero. This makes the calculation of the weighted distances impossible. Therefore, only the Euclidean distance can be used which, however, results in a poor success rate.

The verification problem was experimentally studied in the following way. For each writer, two different classes were defined. The first class (ω_1) contained the genuine samples of the specific writer, whereas the other (ω_2) contained the samples of the remaining 49 writers. Thus, a total of 50 pairs was formed and evaluated. For each pair the individual cluster centers $m_i\,(i = 1, 2)$ and covariance matrices C_i (defined as previously) were evaluated using 2249 out of the 2250 points in the feature space. After that, the remaining point was classified into one of the two classes (writer i or not) based on the minimum weighted distance as (19) indicates. The leave-one-out-method was repeated 2250 times. The type of feature achieving the maximum identification rate (Table 4, 10 segments partition) was used in this experiment, as well. The results are presented in Table 5. The verification error is of the order of 5% for both words when the weighted distance is used. The mean value of R was evaluated for the 50 writers in order to have an approximate measure for the verification rate. However, it is noted that the classification error is much smaller than that determined by the separability measure R, since the orientation of the cluster hyperellipsoids in the feature space, in general, is different from the direction defined by the line $\bar{m}_i - \bar{m}_j$.

5.2. Classification using neural networks

The performance of the neural network classification scheme is investigated in case of the general identification problem using 50 writers. So, the capability of the network to separate the feature space and correctly classify the majority of the samples is tested. The classifier employed was a three-layer neural network with 20 neurons for the input and the hidden layer. The 20-dimensional feature vector (10 level partition was used) was inserted into the input layer. The output layer consists of six neurons. Consequently, the six-bit binary number at the output of the network will point to one of the fifty clusters. The network was trained using half of the samples (1150). The rest samples (1100) were used to test the network in the discrimination procedure. The method was applied to both the English and the Greek word. The classification results obtained by means of this procedure are given in Table 6. The identification error is in the order of 3.5%. It is noted that the identification success is higher than that obtained using the classical Bayesian approach followed previously. This is due to the fact that employing neural networks the feature space is divided into decision regions independently of the underlying cluster statistics. The training procedure terminates when the sum of squared errors becomes quite small as it is shown in Fig. 14. Actually, this error may be relatively large when it is used as a criterion for stopping the training of the network, especially when the number of training samples is small and the clusters are not well separated in the feature space [19]. Therefore, it is preferable to train the network so that the best performance from the available samples is achieved by allowing the error to become as small as possible through several epochs.

The verification issue, which means separate one cluster from the rest 49, can be solved much easier using a simpler neural structure. This happens because the feature space is to be separated into two different regions only. The experimental results acquired were better than those taken when solving the general identification problem. However, for each writer a different network is required. Cumulative experimental results are shown in Table 7. The verification error is of the order of 2%.

6. Conclusions and discussion

A new feature vector is proposed for writer discrimination by means of morphologically transforming the projection function of a word. This waveform is a description of the way the pixels of the word are distributed along the direction of projection. The feature vector is formed ignoring the blanks between the letters since in this case the separability of the clusters is better. Furthermore, the use of trapezoidal windows (segments) for the formation of the feature vector results in Gaussian statistics and higher separability in the feature space. An extensive study for the statistics (pdf) of the feature components was provided using the Kolmogorov–Smirnov fit test. The dimensionality of the feature vector is determined by the length of the word, the number of SEs used to process

Table 4

(a) Identification rate in 50 writers case: English word

Partition level	5	6	7	8	10	12	15
Euclidean distance	48.18	46.70	54.85	56.14	56.74	51.22	44.29
Weighted distance	69.51	71.59	80.77	87.59	92.48	X	X

(b) Identification rate in 50 writers case: Greek word

Partition level	5	6	7	8	10	23	25
Euclidean distance	50.85	55.30	58.56	59.41	63.74	65.22	60.57
Weighted distance	70.85	77.37	83.89	86.11	92.63	X	X

Table 6
Classification results for the general identification problem using the neural network approach: partition level is 10

Language	Classification success
English	1061/1100 (96.5%)
Greek	1069/1100 (97.0%)

Table 5
Classification results for binary decision problem (verification): partition level is 10

Language and similarity measure used	Mean separability measure R	Maximum expected classification error based on R (%)	Experimental classification error (%)
English Euclidean	0.75	22.30	12.25
English weighted	0.75	22.30	6.35
Greek Euclidean	0.79	21.01	14.73
Greek weighted	0.79	21.01	5.57

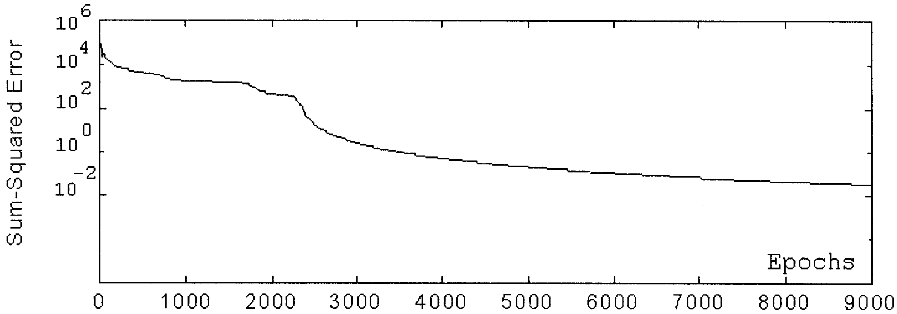

Fig. 14. Sum of squared errors in the identification problem: English word. Partition level is 10.

Table 7
Cumulative experimental results for the verification problem using the neural network approach: partition level is 10

Language	Classification success
English	1075/1100 (97.7%)
Greek	1085/1100 (98.6%)

the obtained waveform as well as by the number of segments the original image is divided into (partition level).

A database was built to test the discrimination capabilities of the proposed feature. Fifty writers were employed for this purpose and two different words of the same length were used, an English word and the corresponding (same meaning) Greek one. It is shown throughout the paper that the identification results obtained using these two words are equivalent, which means that the proposed feature is independent of the language used. The database can be accessed by any researcher though the Internet.

The classification results obtained using the proposed feature can be considered satisfactory given that only one word is employed for writer discrimination. Two different classification schemes were tested, namely, the conventional Bayesian classifier and the neural networks. The verification problem was solved considering the writer to be verified as belonging to class ω_1, while the rest of the writers form the class ω_2. The verification error using the Bayesian approach is in the order of 5% as shown in Table 5. It is shown in Table 7 that using the neural nets this error becomes quite smaller (2%). In the general identification case, where each class corresponds to a different writer, the classification error is in the order of 7% as shown in Table 4. The corresponding error when neural networks are employed is found to be 3.5% (Table 6). All experimental results were obtained using a 20-dimensional feature vector, which corresponds to a 10-level partitioning. For this partitioning the highest identification rate was acquired for the specific length of the words in our data base (Table 4).

Furthermore, it was proved that the components of the proposed feature vector are correlated (Tables 2 and 3). Thus, the intrinsic dimensionality of the feature space which can be obtained using the eigenvalues of the covariance matrices is small (Table 1). In order to avoid a complicated theoretical analysis to test the separability of the clusters in the feature space employing the covariances and the correlation of the features, a separability measure was proposed described by Ref. [14]. This measure was statistically analyzed and used successfully throughout the experiments (Table 5).

The computational load required to carry out the whole procedure can be divided into two parts. The feature extraction step and the classification process. Feature extraction, which is the main objective of this paper, is performed in milliseconds for each of the words in our database, using MATLAB routines (for thresholding, thinning, projecting and resampling) and simple min–max filters for morphological transformations. The computer system used was a Pentium-133 running Windows-NT and MATLAB 5. The time required to perform classification depends on the method employed (Bayesian or Neural) and the training procedure (leave-one-out-method, back-propagation, etc.). The training procedures in our experiments were of the order of minutes.

For further classification improvement more than one word can be used by means of fusion techniques [3].

7. Summary

In this work a writer identification method is proposed, which is based on the use of a single word. A new feature vector is employed by means of morphologically transforming the projection function of the word. First, the image of the word is properly preprocessed (thresholded thinned) and then projected onto the horizontal direction. The obtained projections are segmented in parts which are morphologically processed in order to obtain the required feature vector. The morphological processing is a type of granulometry. Two different types of windows are applied to the segments of the projection to control the flow of information from one part to the other. The blanks between the letters are also taken into consideration in the formation of the feature vector. An extensive study for the statistics (pdf) of the feature components is provided employing the Kolmogorov–Smirnov fit-test.

The identification success depends on the dimensionality of the feature vector which in turn depends on the length of the word, the length of the SEs and the partition level. Furthermore, it is proved that the components of the proposed feature vector are correlated. In order to avoid a complicated theoretical analysis to test the separability of the clusters in the feature space, a separability measure is proposed. This measure is statistically analyzed and used successfully throughout the experiments.

A database was built to test the discrimination capabilities of the proposed feature. Fifty writers were employed for this purpose and two different words of the same length were used, an English word and the corresponding (same meaning) Greek one. It is shown throughout the paper that the identification results obtained using these two words are equivalent, which means that the proposed feature is independent of the language used. The database can be accessed by any researcher through the Internet.

The capability of the proposed feature for writer discrimination is extensively studied. Two different classfication schemes are tested, namely, the Bayesian classifier and the neural networks. In the classification procedure the binary decision problem and the general classification problem are studied. The classification results obtained using the proposed feature can be considered satisfactory given that only one word is used for writer discrimination. The verification error using the Bayesian approach is of the order of 5%, while for the neural nets this error becomes quite smaller (2%). In the general identification case the classification error is of the order of 7%, whereas the corresponding error when neural networks are used is found to be 3.5%.

References

[1] R. Plamondon, G. Lorette, Automatic signature verification and writer identification – The state of the art, Pattern Recognition 22 (2) (1989) 107–131.

[2] M. Yoshimura, F. Kimura, On a formal measurement of individual variability appeared in handwritten characters, Trans. IECE Japan 9 (1980) 795–802.

[3] E.N. Zois, V. Anastassopoulos, Fusion of correlated decisions for writer identification, Proceedings of the IEE Third European Conference on Handwriting Analysis and Recognition, Brussels, 14–15 July 1998.

[4] V. Clement, K. Steinke, R. Naske, The application of image proccesing and writer recognition techniques to the forensic analysis of handwriting, Proceedings of the International Conference Secur., West Berlin, 1980, pp. 5–11.

[5] V. Clement, Forensic writer recognition, Proceedings of the NATO Advanced Studies Institute, 1981, pp. 519–524.

[6] K. Steinke, Recognition of writer by handwriting images, Pattern Recognition 14 (1981) 357–364.

[7] E.N. Zois, V. Anastassopoulos, Methods for writer identification, Proceedings of the International Conference on Circuits and Systems, Rodos Island, Greece, 13–16 October 1996, pp. 740–743.

[8] F. Mihelic, N. Pavesic, L. Gyergyek, Recognition of writer of handwritten texts, Proceedings of the International Conference on Crime Countermeasures, Kentucky University, Lexington, 1977, pp. 237–240.

[9] M. Levine, Vision in Man and Machine, McGraw-Hill, New York, 1985.

[10] T. Pavlidis, Algorithms for shape analysis of contours and waveforms, IEEE Trans. Pattern Anal. Machine Intell. PAMI-2 (1980) (4) 301–312.

[11] Y. Qi, B.R. Hunt, Signature verification using global and grid features, Pattern Recognition 27 (12) (1994) 1621–1629.

[12] R. Nagel, A. Rosenfeld, Computer detection of freehand forgeries, IEEE Trans. Comput. C-26 (9) (1977) 895–904.

[13] S.W. Lee, J.S. Park, Nonlinear shape normalization methods for the recognition of large set handwritten characters, Pattern Recognition 27 (7) (1994) 902–985.

[14] T. Hilderbrand, W. Liu, Optical recognition of handwritten Chinese characters: advances since 1980, Pattern Recognition 26 (2) (1993) 205–225.

[15] S.R. Sternberg, Grayscale morphology, Comput. Vision Graphics Image Process. 35 (1986) 333–355.

[16] P. Maragos, Pattern spectrum and multiscale shape representation, IEEE Trans. Pattern Anal. Machine Intell. PAMI-11 (7) (1989) 701–716.

[17] V. Anastassopoulos, A.N. Venetsanopoulos, The classification properties of the spectrum and its use for pattern identification, Circuits Systems Signal Process. 10 (3) (1991) 117–143.

[18] E.N. Zois, V. Anastassopoulos, 'writer identification data base II', ftp://anemos.physics.upatras.gr/pub/handwriting/HIFCD2, Electronics Laboratory, Physics Dept., University of Patras, Patras, Greece, 1998.

[19] J.P. Drouhard, R. Sabourin, M. Godbout, A neural network approach to off-line signature verification using directional pdf, Pattern Recognition 29 (3) (1996) 415–424.

[20] A. Papoulis, Probability, Random Variables and Stochastic Processes, McGraw-Hill, New York, 1994.

[21] W.H. Press et al., Numerical Recipes, Cambridge University Press, Cambridge, 1992.

[22] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1970.

[23] K. Fukunaga, Statistical Pattern Recognition, Academic Press, San Diego, 1990.

[24] M. Abramowitz, I.E. Stegun, Handbook of Mathematical Functions, U.S Dept. of Commerce, National Bureau of Standards, 1972.

About the Author—ELIAS ZOIS was born in Athens, Greece in 1971. In 1994 he received his B.Sc. degree in Physics from Physics Department, University of Patras, Greece. In 1997 he received his M.Sc. in Electronics from Electronics Laboratory, University of Patras, Greece. His Master Thesis was on handwritten text analysis and writer identification. He is now a research fellow in the same laboratory working towards his Ph.D. His main interests are analog circuits design, digital image processing and pattern recognition and classification with emphasis on handwritten text analysis and discrimination.

About the Author—VASSILIS ANASTASSOPOULOS was born in Patras, Greece, in 1958. He received the B.Sc. degree in Physics in 1980 from the University of Patras, Greece and the Ph.D. in Electronics in 1986 from the same University. His Ph.D. Thesis was on Digital Signal Processing and in particular, on Delta Modulation Filters.

From 1980 to 1985 he was employed as a research assistant in the Electronics Laboratory, University of Patras. From 1985 to 1989 he was a lecturer in the same Laboratory. From 1989 to 1990 he worked as a research associate in the Department of Electrical Engineering, University of Toronto, on Nonlinear Filters and Pattern Recognition and Classification Techniques. From 1992 he is an Assistant Professor in Electronics Laboratory, University of Patras, Greece. Since 1990, he has been in close cooperation with AUG Signals Ltd, a Canadian company, working on Radar Signal Detection, IR Image Processing and Data Fusion. He worked with this company during his sabbatical in Canada (1994–1995).

His research interests are within the scope of Digital Signal Processing, Image Processing, Radar Signal Processing and Pattern Recognition and Classification. He is a member of the IEEE.